Хетил Ро Хауге
(Университет в Осло)

КАК СЕ ПРАВИ РЕЧНИК ОТ ЕДИН МАЛЪК ЕЗИК ЗА ЕДИН МАЛКО ПО-ГОЛЯМ ЕЗИК ПОЧТИ БЕЗ СРЕДСТВА
(норвежко-български речник)

Abstract: The article reviews tools and general conditions for low-budget lexicography, from data collection to print, outside professional lexicographic institutions. It ranges from tools available at the end of the 1990 s up to the present day, with focus shifting from strictly one-platform solutions at the start to multi-platform tools at the present time.

Keywords: lexicography, bilingual dictionaries, XML, XSLT

Към първата част от заглавието може да се добави и „за доста дълго време“, защото работата вече премина границата на две столетия и още не е завършена... Ключът към разбирането на този факт лежи във фразата „почти без средства“ – проектът се извършва от ентусиасти и е имало паузи от няколко години поради промени в заетостта на участниците.

Инициативата е подета през 1998 г. от български езиковед – скандинавист, който предложи да се направи норвежко-български речник с малък или среден по размер словник. В годините след това неколцина други ентусиасти вземат участие в работата (някои от тях продължават), като авторът на тази статия е участвал като коректор на обработените файлове – главно като програмист и технически редактор. През това време настъпват промени в развитието на информационната технология, поради което част от взетите решения стават неприложими в бъдеще. В тази статия ще се спрем на някои от тях.

Поради оскъдността на средства и време началната задача беше да се намери готов словник: секцията „Лексикография“ при Катедрата за лингвистика и скандинавистика при Университета в Осло ни предостави безвъзмездно електронната база на *Bokmålsordboka 1993*, еднотомен тълковен речник на „букмол“, най-разпространеният от двата писмени варианта на норвежкия език. Беше ни предоставен не само словникът заедно с тълковните дефиниции, а и граматическата анотация и фразеологията. Базата от лексикални данни съществуваше в два формата: текст с кодирания, базирани върху неазбучни знаци (#, >, $, @, ...), които съответстват на съставните елементи на речниковата статия, и текст във формат SGML. Тъй като последният вариант беше твърде сложен (а „по-простият“ му наследник XML беше по това време съвсем нов), избрахме варианта с неазбучните кодирания, за който предлагаме един пример:

```
NB001 abbed
NB001a M1
ARTNR 13
TR007
..OPP #>$Cabbed@ m1
..ETY (lat. $Babbas@, opph arameisk $Babba@ ‚far; munk', jf norr $Babb$1ati@)
..DEF leder for et munkekloster
=
NB001 abbedi
NB001a N5
ARTNR 14
TR007
..OPP $Cabbedi>@ n3
..ETY (gj lty fra mlat)
..DEF kloster som blir ledet av en abbed
=
```

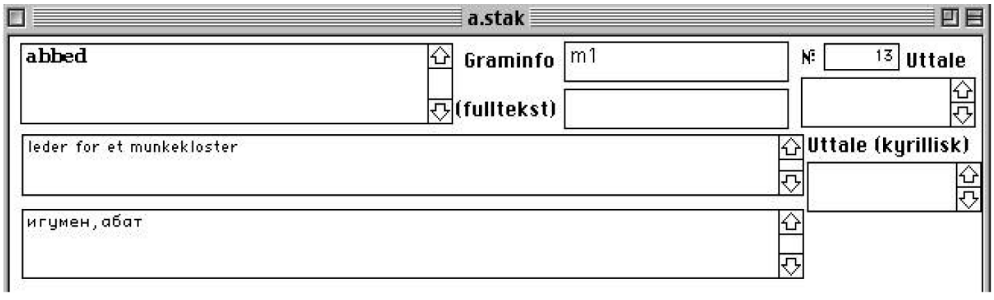
Показани са две речникови статии. Това, което ни интересуваше, бяха заглавните думи (..OPP; *oppslagsord* ‘заглавна дума’): abbed ‘игумен, абат’ и abbedi> ‘манастир, абатство’, заедно с кодовете за род и родов разред (m1, n3) и норвежкото тълкуване (..DEF; *definisjon* ‘тълкуване’). Последното не е част от новия речник, а е само в помощ на авторите на речника, за да намерят най-подходящите български преводи. При по-сложни статии има, разбира се, по няколко значения и/или фразеология към заглавната дума. Важно е означаването на ударени гласни, например „i>“ в abbedi> – ударението в повечето домашни норвежки думи и в много чужди пада върху първата сричка и затова в този речник, ако ударението не е отбелязано, това означава по подразбиране ударение на първата сричка.

Особена трудност представляваше от белязването с тилда в оригиналното кодиране, тъй като съставните думи трябваше да бъдат представени като отделни статии:

```
NB001 alarm
NB001a M1
ARTNR 681
TR007
..OPP #>$Calar>m@ m1
=
NB001 alarmapparat
ARTNR 682
TR007
..OPP $C~apparat@
=
```

В горния пример тилдата при втория код ..OPP трябваше да се възстанови заедно с форматираното съдържание на първия код ..OPP: *alar‘m*, за да може да се получи втората заглавна дума *alar‘mapparat.* Материалът от кода NB001 не можеше да бъде използван, тъй като е предназначен само за горните колонтитули и не съдържа информация за ударението.

Следващата стъпка беше избирането на лексикографска среда за работа. Техническият редактор реши, че въпреки че изтъкнати лексикографи създават големи речници в текстообработващи програми (например Word) и даже успяват да спазват унифицирано структуриране на речниковите статии, за този проект е необходим по-строг режим на работа, който да налага стриктно спазване на приетата структура. Избраната среда за работа беше програмата HyperCard на Apple – програма, която съчетава бързо търсене в база от данни и визуално представяне на данните, подредени в отделни графи, отговарящи в нашия случай на структурни елементи на речниковата статия (виж Фигура 1 по-долу).



Фиг. 1. Представяне на данните в HyperCard

Извличането на нужните данни от лексикалната база беше направено със скриптовия език Applescript. Във времето преди Unicode кирилицата беше кодирана с AppleCyrillic. Отделните автори нямаха възможност да видят резултатите от създаването и редактирането на речниковите статии в графично оформен вид като страница от речник. Техническият редактор извършваше две стъпки: конвертиране на изходния формат посредством вложените възможности в HyperCard в текстовия формат MIF (Maker Interchange Format, специфичен за издателската система FrameMaker и далечен родственик на HTML), който се чете от FrameMaker и позволява да се създаде PDF файл. Получените PDF-и се предоставяха на сътрудниците на речника. Това усложнение се дължеше на вече споменатото условие „почти без средства“ – само техническият редактор разполагаше с тази (относително скъпа) програма.

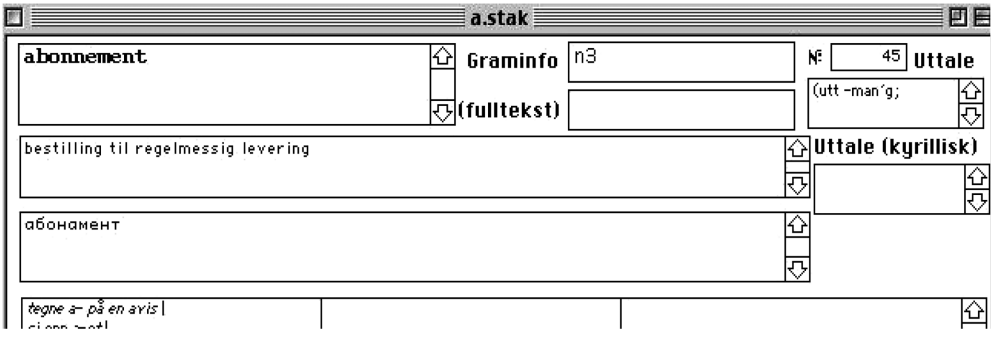
След 2004 г. HyperCard вече не се поддържаше и от Apple и старата ни технология стана невъзможна за употреба. Частично решение беше средата

за програмиране Runtime Revolution (сега преименувана в LiveCard), нещо като наследник на HyperCard, с разширен вариант на езика за програмиране и с възможност да се генерират версии за различни платформи.

По това време вече от няколко години беше разпространен Unicode, но Runtime Revolution трудно се справяше с него и бяхме принудени да прибегнем до транслитериране на кирилските символи. Това не беше голям проблем, защото с AppleScript транслитерираният текст можеше да се конвертира в стандартна кирилица.

На този етап материалът се състоеше от файлове с обем около три четвърти от крайната цел; най-новите във формат Runtime Revolution/LiveCard, съвместим донякъде, но не напълно, с формата на HyperCard. Главният проблем беше, че програмата FrameMaker, която преди се използваше за генериране на типографски оформени страници, вече не съществуваше за платформата Макинтош. Техническият редактор прецени, че XML е единственото решение, което ще позволи пълноценна бъдеща работа (тук може да се подчертае погрешното първоначално решение да се избере неазбучно кодиране вместо SGML). Аргументите в полза на XML бяха няколко: за всички по-големи платформи има безплатни или не много скъпи програми за редактиране на XML; с помощта на XSL трансформации отделните сътрудници могат с XML редактор бързо да генерират графично оформена уеб страница от файла, над който работят; конвертирането от „картовия“ формат към XML е сравнително просто, понеже графите в HyperCard съответстват точно на елементите в XML. Данните във Фигура 2 по-долу могат да се представят в XML фрагмент, както следва:

```
<headword>abonnement</headword>
<pronounce>-man‘g</pronounce>
<graminfo>n3</graminfo>
<bgtransl>абонамент</bgtransl>
```



Фиг. 2. Данни от речниковата статия

В началото намаляването на словника приблизително наполовина (от около 65 хиляди думи в изходния тълковен речник) се правеше ръчно, като в качеството на ръководство (при липса на друго) се използваше друг двуезичен речник с приблизително същите размери като желаните от нас. Друго предимство от развитието на информационните технологии беше, че след като получихме от Катедрата по лингвистика и скандинавистика списък на 50 000-те най-често срещани леми в норвежкия език, въведохме в XML версията при всички заглавни думи/леми анотация, показваща ранга на лемата в този списък. Изключването на леми продължава да е ръчно по преценка на сътрудниците, като се вземат пред вид особеностите на двуезичния речник в сравнение с тълковния.

Все още има какво да се прави преди окончателната предпечатна обработка на речника, но тя по всяка вероятност ще стане чрез импортиране на XML файловете в издателската система InDesign. При този процес всеки отделен XML елемент ще има свое съответствие в даден стил за абзац и текст в InDesign, като стиловете ще са предварително форматирани за нуждите на книжното тяло. Безплатно решение би било конвертиране с XSL-FO трансформации във Formatting Objects, т.е. вид XML, в който елементите не описват функции като заглавна дума, граматическа анотация, превод и др., а типографски части от печатната страница. XML-ът може да се конвертира в PDF с помощта на добър XML редактор, платен или безплатен. Има и решение с Word и малко повече ръчна работа: след селектиране се копира уеб страницата, получена от XSL трансформация (вж. горе), и се вмъква в Word. При тази операция се запазва в общи линии форматирането, макар че се губи предимството на работа със стилове за абзаци и текст. Word разполага с възможност за създаване на активни заглавни колонтитули, но за да се използва, е необходимо заглавните думи да се форматират уникално, което да позволява превръщане в стил посредством търсене и замяна.

На фона на това сравнително сполучливо и почти безпроблемно прехвърляне от един забравен от историята формат в XML няма да е излишно да хвърлим поглед на друг все още съществуващ формат и възможностите за конвертиране в XML. Техническият редактор участва в още един проект за речник: норвежко-литовски речник. За този проект получихме безвъзмездно словника от Берков и др. 2003, *Большой норвежско-русский словарь*, в Word. Съществува и TeX за предпечатна обработка, но тази версия е собственост на издателството и не ни беше предоставена. Проблемите, свързани с конвертирането на този речник, са описани в Хауге & Берг-Улсен 2005 и тук стига да се спомене, че форматирането (шрифт, курсив, ...) е база за приписването на XML елементи, но броят на потенциални елементи е по-голям от броя на комбинациите за форматиране и затова трябва да се вземе предвид и мястото на елемента в речниковата статия, което усложнява задачата. След публикуването на статията през 2005 г. се е появил новият формат .docx, който в основата си е XML, а потребителят го „вижда“ като типографично форматиране. Този

формат позволява експортиране в много сложен XML, труден за по-нататъшно обработване, понеже трябва да поддържа всички възможности за форматиране на Word. По-добра алтернатива за получаване на XML от формата .docx е да се работи с друга програма, която има опция за експортиране във формат XML. TextEdit за Макинтош дава по-прости резултати от Word, а най-добрите са на OpenOffice (или NeoOffice за Макинтош), с точно превеждане на форматирането в XML елементи, които понякога са повече, отколкото са необходими, но лесно може да бъдат сведени до минималния необходим брой.

ЛИТЕРАТУРА

Берков и др. 2003: Berkov, V., H. Haraldsson, St. Kottum. *Stor norsk-russisk ordbok // Большой норвежско-русский словарь*. Oslo: Kunnskapsforlaget.

Хауге & Берг-Улсен 2005: Ро Хауге, Х., Ст. Берг-Улсен. Конвертиране на речник от Word в XML: норвежко-литовски речник. // *Лексикографски преглед*, № 8, с. 41–46.

Bokmålsordboka 1993: *Bokmålsordboka. Definisjons- og rettskrivningsordbok.* Oslo: Kunnskapsforlaget.